D310–D314 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj099

SCOPPI: a structural classification of protein–protein interfaces

Christof Winter[1], Andreas Henschel[1], Wan Kyu Kim[1] and Michael Schroeder[1,*]

[1]Biotechnological Centre of TU Dresden, Tatzberg 47-51, 01307 Dresden, Germany

Received August 15, 2005; Revised and Accepted October 17, 2005

ABSTRACT

SCOPPI, the structural classification of protein–protein interfaces, is a comprehensive database that classifies and annotates domain interactions derived from all known protein structures. SCOPPI applies SCOP domain definitions and a distance criterion to determine inter-domain interfaces. Using a novel method based on multiple sequence and structural alignments of SCOP families, SCOPPI presents a comprehensive geometrical classification of domain interfaces. Various interface characteristics such as number, type and position of interacting amino acids, conservation, interface size, and permanent or transient nature of the interaction are further provided. Proteins in SCOPPI are annotated with Gene Ontology terms, and the ontology can be used to quickly browse SCOPPI. Screenshots are available for every interface and its participating domains. Here, we describe contents and features of the web-based user interface as well as the underlying methods used to generate SCOPPI's data. In addition, we present a number of examples where SCOPPI becomes a useful tool to analyze viral mimicry of human interface binding sites, gene fusion events, conservation of interface residues and diversity of interface localizations. SCOPPI is available at http://www.scoppi.org.

INTRODUCTION

Understanding protein interactions bears the key to understand many cellular processes. The function of newly discovered proteins can often be inferred by identifying its interaction partners, and many human diseases can be traced to aberrant protein–protein interactions (1). Besides efforts for large-scale interactome maps (2–8), protein interactions and their interfaces have also been studied intensively on the structural level (9–13). Although various interaction databases and prediction servers exist [such as 3DID, PIBASE, STRING, IntAct, ProMate, InterPreTS or PSIMAP (14–21)], few focus on the geometrical aspects of the domain–domain association.

SCOPPI, the structural classification of protein–protein interfaces, contains domain–domain interactions of proteins with known structure. The domain interaction criterion is distance-based and follows the approach of PSIMAP as described previously (22,23). Domains are defined according to SCOP (24). Unlike other domain interaction databases [such as 3DID (14) or PIBASE (15)], SCOPPI does not focus on the resulting domain interaction network, but rather on the domain sequences and interacting residues that form the interface. To this end, SCOPPI provides multiple sequence alignments of all members within a SCOP family. The combination of multiple sequence and structural alignment of SCOP family members allows for a thorough classification of binding sites to other domains: For all aligned domains in a SCOP family, binding sites to other domains called *faces* that are overlapping according to sequential and structural features are clustered into distinct *face types* (25). One SCOP family can have many face types, depending on the binding site diversity of its family members. Two interacting face types constitute an *interface type*. By means of our method, ~8400 interface types are identified.

SCOPPI can be queried by SCOP family, superfamily, PDB IDs or keywords. Results can be accessed in different views (multiple sequence alignments with highlighted interface residues, screenshots of domains and the interface) and filtered by sequence redundancy and interface size.

In the Examples section, we will demonstrate how SCOPPI's features are useful for a variety of questions.

SCOPPI: IDEA, USAGE AND FEATURES

The idea of SCOPPI is to investigate domain–domain interfaces in proteins of known structure. Domains are identified according to the SCOP classification of protein structures (24). In accord with other interface definitions (26), we define two domains to interact if they have at least five residue–residue contacts within 5 Å (21). SCOPPI presents a multiple sequence alignment of domains within a SCOP family.

*To whom correspondence should be addressed. Tel: +49 351 463 40062; Fax: +49 351 463 40061; Email: ms@biotec.tu-dresden.de

Interface residues are highlighted in the aligned sequences, which affords a geometrical classification of binding sites, the unique feature of SCOPPI. Various other interface characteristics such as permanent or transient nature of the interaction, interface size (the loss of accessible surface area upon complexation, ΔASA), domains on the same or different polypeptide chains and number of interacting atoms and residues are available.

Query options

SCOPPI can be queried for a SCOP family, superfamily, one or several PDB identifiers or a keyword. For keywords entered, SCOP family and superfamily description, PDB headers and InterPro abstracts (27) are searched. SCOPPI can be further browsed by SCOP family descriptions alphabetically and by the Gene Ontology hierarchy (28). All queries will finally result in the display of sequences or screenshots of interacting domains along with interface characteristics.

Data view

A typical query result is presented in Figure 1. The data are organized in a table: each row represents one domain–domain interaction, and each column depicts one property of this interaction. In the default view, SCOPPI shows resolution and 4-letter code of the source file, sequences of the two domains in full length with highlighted interacting residues, the *face types* for both domains, the interaction type and a link to GoPubMed (29). Both sequence columns are grouped by SCOP family. If SCOPPI was queried for a family and not for a PDB identifier, this family always appears on the left. The 'View' selector to the very left above the result table is used to obtain different views on the data: to access all interaction properties including the SCOP unique identifier for each domain, it can be changed to 'All'. 'Structures' presents a view without sequences, but with screenshots for each interaction. Images are clickable to obtain a larger version.

Display of sequences

Interacting residues are displayed in upper case letters, non-interacting in lower case. The default coloring highlights interacting residues for better identification. It can be changed with the 'Color' selector to assign different colors to different face types. A simple conservation overview is provided by assigning a color to each residue depending of the frequency in the column of the multiple sequence alignment. Residues can be further colored by physicochemical type. 'Sequence display' facilitates switching between aligned sequences, raw sequences without gaps and only the aligned interfaces, where three dots indicate four or more left out non-interacting residues.

Filter options

Since lots of identical sequences exist among SCOPPI's over 90 000 domain interactions, we provide non-redundant sets at various sequence identity levels. A default 90% cut-off leads to ∼15 000 different domain–domain contacts. Non-redundant sets are available via the 'Redundancy cut-off' selector. To filter out small interfaces, we calculate the change in accessible surface area ΔASA and provide an interface size cut-off selector for 600, 1400 and 2000 Å^2.

(a)

(b)

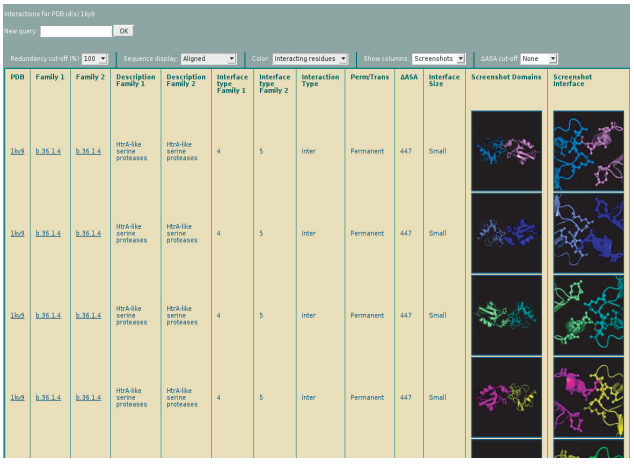

Figure 1. (**a**) A typical query result. SCOPPI displays sequences of interacting domain pairs, aligned by families. Each row represents one interaction, while columns describe various aspects of that interaction. Views on the data, filters and coloring can be changed. Here, residues are colored by conservation within the family. (**b**) When switching to the structure view, SCOPPI shows screenshots of two interacting domains (left) and their interface (right). Various interface characteristics such as size of interface, number of involved residues, permanent or transient nature of the interaction are further available. Explanations for these characteristics appear as the mouse is moved over the column headers and are listed on the help page.

EXAMPLES

SCOPPI provides a multiple sequence alignment of domains and their interacting residues within a family. The multiple alignment of these interaction interfaces in combination with the clustering and classification of binding sites is a unique feature of SCOPPI. The following examples will illustrate how this feature can provide interesting insight into various fields.

We use the term *face* for a binding site on a single domain and *interface* for the interacting faces of two domains. *Face type* defines similar binding sites for members in a SCOP family, and the combination of two face types forms an *interface type*.

An elaboration of these examples along with meaningful screenshots is presented online at SCOPPI's help page. In the

following, IDs such as 1kim refer to PDB structures (30), and a.48.1.1 to SCOP families (24).

Binding site similarity: viruses mimic interfaces

Chemokines play a key role in leukocyte recruitment and migration. A query for the term 'chemokine' in SCOPPI finds, among others, the Interleukin-8 (IL-8) like chemokines family (d.9.1.1). Following the family link in the results lists all interactions of chemokine family members with other domains: SCOPPI shows that IL-8 like chemokine domains can associate with domains of the same family, forming homodimers, or that they can associate with members of the viral chemokine binding protein M3 family (b.116.1.1). Further coloring by *face type* using the color selector reveals that the viral protein binds to the same face type of the chemokine that is used for homodimerisation. SCOPPI also displays this information in the 'Face type Family 1' column. Alexander *et al*. (31) report that viral protein M3 indeed employs structural mimicry to sequestrate chemokines.

Binding site diversity: where do cytokines bind to their receptors?

Some families' members display a considerable variety of face types when interacting with members of another family. Querying SCOPPI for the long-chain cytokines (a.26.1.1) reveals such an example. Domains of this family appear as part of human cytokines, human growth hormone and prolactin. When interacting with their receptors, the ligand's cytokine domain binds to the fibronectin type III domain of the receptor. SCOPPI reports that there are 10 different interface types for this interaction ('Face type Family 1' column on the right). To confirm this, we superimpose 3D interactions of several examples structurally aligning the cytokine domain (Figure 2). The cytokine domains are shown in black, with the associated fibronectin domains in various colors. There are clearly numerous different interaction sites (*face types*) on the cytokine domain surface.

Binding site conservation: how well conserved is the trypsin pocket?

Consider trypsin-like serine proteases that are found in the family of eukaryotic proteases (b.47.1.2). The active site of these enzymes is formed by a catalytic triad of three residues: histidine, aspartic acid and serine (in sequential order). Owing to the obvious importance of these three residues, we expect them to be conserved throughout all members of the serine protease family. To verify this, the user may enter b.47.1.2 in SCOPPI, apply a redundancy level of 50% for a better overview and select 'Conservation' from the color selector. The conservation percentage is simply calculated by counting the number of residues of the same type in that column divided by all residues in that column. Residues with a value above 90% will display in red, those with a value below 10% in purple. A color legend pop-up is available through a hyperlink next to the selectors. For the serine proteases, SCOPPI reveals a highly conserved region AAHC with the catalytic histidine residue. Asp (D) is also well conserved (DIxLxxL motif). Serine (S) is found inside a conserved GDSGGP motif. It is striking, however, that the serine is not fully conserved—between 10 and 20% of the family members are missing

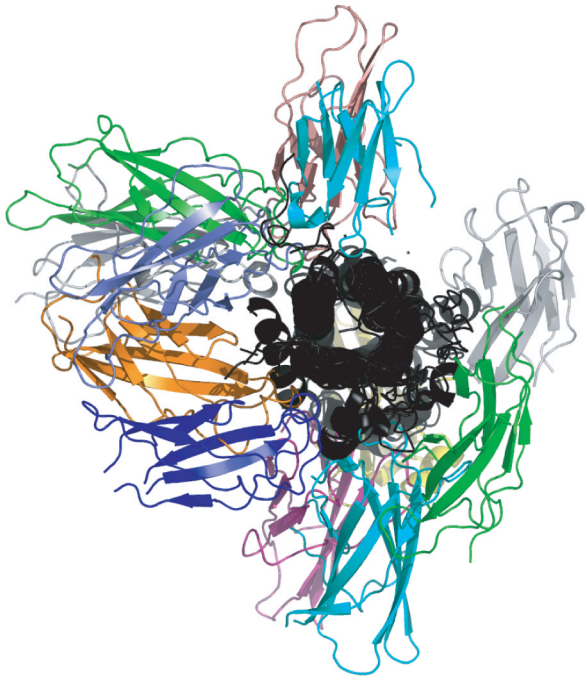

Figure 2. Binding site diversity: superimposed examples of an interaction between a cytokine domain (black) and a fibronectin domain (various colors). The examples were structurally aligned to the cytokine domain. According to SCOPPI, a total of 10 different *face types* are identified on the cytokine domain surface for this interaction.

the serine at this position. Closer examination reveals that in these cases serine had been subject to site-directed mutagenesis studies and was changed to alanine.

Binding site orientation: gene fusion

The interface type classification of SCOPPI defines groups of pairs of domains that associate in the same geometric orientation (i.e. clusters of domain pairs with the same *interface type*). In addition, SCOPPI provides the information if two interacting domains reside on the same or on separate polypeptide chains (i.e. if the interaction type is *intra* or *inter*). If two domain pairs interact in the same orientation, but the interaction type is *intra* in one and *inter* in the other case, the reason behind this observation might be a gene fusion event.

Such an example is found among domains of the c.1.2.1 family (Histidine biosynthesis enzymes) interacting with domains of family c.23.16.1 (Class I glutamine amidotransferases). SCOPPI's face type column on the right informs that there are *inter* and *intra* cases displaying the same two face types. The matching face types can also be identified just by looking at the highlighted interacting residues of the aligned sequences.

Taking two PDB files of the listed cases, 1gpw and 1ox4, and displaying the interacting domains confirms the above finding at structural level: both PDB files describe the crystal structures of Imidazole Glycerolphosphate Synthase, which catalyzes formation of the imidazole ring in histidine biosynthesis. The functional enzyme consists of a glutamine amidotransferase domain and a cyclase domain. In Thermotoga maritima, a hyperthermophile bacterium, these domains are

located on two separate polypeptide chains, forming a heterodimeric protein. In yeast, the two domains are fused together, as it is common in plants and fungi (32). SCOPPI nicely picks up this example by its classification of geometrically distinct interface types. In total, we identify 59 of such examples.

MATERIALS AND METHODS

The contents of SCOPPI are results of the followings steps: 17 083 protein structures are taken from the Protein Quaternary Structure Server PQS at EBI (http://pqs.ebi.ac.uk) (33), which offers coordinates for likely quaternary states of structures contained in the PDB (30). These correspond to 14 061 structures in PDB (30). Applying SCOP 1.67 domain definitions (24) yields 70 527 domains grouped into 2209 families. Interacting domain pairs within one PQS file are determined using PSIMAP (20,22). PSIMAP considers a pair of domains as interacting if at least five interacting residue-residue contacts exist within a 5 Å distance. We define the interacting residues of each domain as *face* and the corresponding pair of faces as *interface*. Furthermore, we distinguish between two interaction modes: *inter* interactions occur between two domains that have different chain identifiers in the PQS entry, whereas *intra* interactions involve domains on the same chain.

Domain sequences are parsed from PQS files. For each family, a multiple sequence alignment is built by MUSCLE (34). The *face* residues are mapped onto the aligned sequences. The clustering of faces is a two-step procedure: first, the aligned sequences are converted to an interface tag (IFT) by representing interface residues by 1s and other residues by 0s, resulting in a linear vector such as 0–000101110—100. The IFTs of each family are clustered into groups with similar patterns. The distance between two vectors u, v is measured by the cosine angle distance $D_{IFT}(u,v) = 1 - \frac{uv}{|u||v|}$, ignoring positions containing gaps. D_{IFT} becomes 0 between identical IFT pairs and 1 between IFT pairs without any common interface residue. Since only faces of highly similar IFT patterns ($D_{IFT} < 0.2$) are grouped together, the resulting clusters consist mostly of equivalent surfaces.

In a second step, geometrical features measuring the similarity between two faces are used to further merge the clusters: upon structural superposition of two domains with MultiProt (35), (i) the overlap of faces—i.e. the percentage of atoms that are within 3 Å of the other face—and (ii) the angle between two lines connecting the domains' common center of mass and the centers of mass of the two faces are calculated. The clustering thresholds for (i) and (ii) were set to 60 and 40%, respectively, which proved best on systematic benchmarking. After this step, the 92 979 domain contacts of SCOPPI are clustered into 8381 distinct interfaces.

A series of non-redundant interface sets are provided at different sequence identity levels ranging from 50 to 100%. To this end, representative sequences for each threshold are generated for each family using CD-HIT (36). Further collating of domain pairs with same face types leads to non-redundant interface sets. For each interaction, the change in accessible surface area, ΔASA is calculated with Naccess (http://wolf.bms.umist.ac.uk/naccess), which is an implementation of the Lee and Richards probe method (37). Conservation coloring: for each family, the number of occurrences of a particular residue divided by the length of the column is calculated for each column of the multiple sequence alignment and then assigned to every residue of that type. If all residues in a column are glycines, every glycine gets a score of 1 therefore, and if 30% were alanines, glycines would get 0.7 and alanines 0.3. Ten rainbow color shades ranging from bright red ($\geq$0.9) to purple (<0.1) display the assigned score. Please note that conservation scores are calculated based on the currently displayed sequences, so they will change with different redundancy levels or ΔASA cut-offs. Please also note that conservation coloring is not meaningful for single PDB files.

GO annotations for PDB files are made available by the GOA project (38). For each family, we provide a simple GO annotation overview by counting GO terms of the PDB files of that family. For the GO categories cellular component (C), molecular function (F) and biological process (P), the three most frequent GO terms are shown. The Medline link of GOA is used to link to the GoPubMed server (29), where the relevant literature for a particular PDB entry can be viewed hierarchically indexed by Gene Ontology terms.

SUMMARY

We present SCOPPI, the structural classification of protein–protein interfaces. SCOPPI is based on structures from the PDB (30) and SCOP domain definitions (24). For each SCOP family, we provide multiple sequence alignments with highlighted interface residues. Our unique sequence- and structure-based classification of binding sites as well as screenshots of the interacting domains and the interfaces allows for a quick examination of the binding geometry between two domains (Figure 1). Additionally, GO annotations and various interface characteristics are available. SCOPPI can be queried by SCOP family, superfamily, PDB IDs, keywords and browsed by family names or the Gene Ontology. Results can be filtered by sequence redundancy and interface size.

The usefulness of these features is illustrated by examples where SCOPPI is found valuable to study various aspects of protein interactions. In particular, these include similarity, diversity, conservation and orientation of binding sites.

ACKNOWLEDGEMENTS

Funding by EFRE project CODI no. 4212/04-07 is kindly acknowledged. Many thanks to Gihan Dawelbait, Andreas Doms, Bingding Huang and Samatha Kottha for contributing valuable data. We further thank Maxim Shatsky for his assistance on using MultiProt (35).

Conflict of interest statement. None declared.

REFERENCES

1. Ryan,D. and Matthews,J. (2005) Protein–protein interactions in human disease. *Curr. Opin. Struct. Biol.*, **15**, 441–446.
2. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

3. Rain,J.C., Selig,L., De Reuse,H., Battaglia,V., Reverdy,C., Simon,S., Lenzen,G., Petel,F., Wojcik,J., Schachter,V. *et al.* (2001) The protein–protein interaction map of *Helicobacter pylori*. *Nature*, **409**, 211–215.
4. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
5. Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
6. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
7. Giot,L., Bader,J.S., Brouwer,C., Chaudhuri,A., Kuang,B., Li,Y., Hao,Y.L., Ooi,C.E., Godwin,B. and Vitols,E. (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.
8. Li,S., Armstrong,C.M., Bertin,N., Ge,H., Milstein,S., Boxem,M., Vidalain,P.O., Han,J.D., Chesneau,A., Hao,T. *et al.* (2004) A map of the interactome network of the metazoan *Celegans*. *Science*, **303**, 540–543.
9. Hubbard,S.J. and Argos,P. (1994) Cavities and packing at protein interfaces. *Protein Sci.*, **3**, 2194–2206.
10. Jones,S. and Thornton,J.M. (1996) Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
11. Valdar,W.S. and Thornton,J.M. (2001) Protein–protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, **42**, 108–124.
12. Gao,Y., Wang,R. and Lai,L. (2004) Structure-based method for analyzing protein-protein interfaces. *J. Mol. Model (Online)*, **10**, 44–54.
13. Bahadur,R.P., Chakrabarti,P., Rodier,F. and Janin,J. (2004) A dissection of specific and non-specific protein–protein interfaces. *J. Mol. Biol.*, **336**, 943–955.
14. Stein,A., Russell,R.B. and Aloy,P. (Jan, 2005) 3DID: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res.*, **33**, D413–D417.
15. Davis,F. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.
16. vonMering,C., Jensen,L.J., Snel,B., Hooper,S.D., Krupp,M., Foglierini,M., Jouffre,N., Huynen,M.A. and Bork,P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, **33**, D433–D437.
17. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A., Margalit,H., Armstrong,J., Bairoch,A., Cesareni,G., Sherman,D. and Apweiler,R. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, **32**, 452–455.
18. Neuvirth,H., Raz,R. and Schreiber,G. (2004) ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.*, **338**, 181–199.
19. Aloy,P. and Russell,R.B. (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics*, **19**, 161–162.
20. Park,J., Lappe,M. and Teichmann,S.A. (2001) Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.*, **307**, 929–938.
21. Gong,S., Yoon,G., Insoo,J., Bolser,D., Dafas,P., Schroeder,M., Choi,H., Cho,Y., Han,K., Lee,S., Choi,H., Oh,D., Lappe,M., Holm,L., Kim,S. and Bhak,J. (2005) PSIbase: the Database of the protein structural interactome MAP. *Bioinformatics*, **21**, 2541–2543.
22. Dafas,P., Bolser,D., Gomoluch,J., Park,J. and Schroeder,M. (2004) Using convex hulls to extract interaction interfaces from known structures. *Bioinformatics*, **20**, 1486–1490.
23. Kim,W.K., Bolser,D.M. and Park,J.H. (2004) Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics*, **20**, 1138–1150.
24. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
25. Kim,W.K. and Ison,J.C. (2005) Survey of the geometric association of domain–domain interface. *Proteins*, **61**, 1075–1088.
26. Tsai,C.J., Lin,S.L., Wolfson,H.J. and Nussinov,R. (1996) A dataset of protein–protein interfaces generated with a sequence-order-independent comparison technique. *J. Mol. Biol.*, **260**, 604–620.
27. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al.* (Jan, 2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
28. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
29. Doms,A. and Schroeder,M. (2005) GoPubMed: exploring pubMed with the gene ontology. *Nucleic Acids Res.*, **33**, W783–W786.
30. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
31. Alexander,J.M., Nelson,C.A., vanBerkel,V., Lau,E.K., Studts,J.M., Brett,T.J., Speck,S.H., Handel,T.M., Virgin,H.W. and Fremont,D.H. (2002) Structural basis of chemokine sequestration by a herpesvirus decoy receptor. *Cell*, **111**, 343–356.
32. Chaudhuri,B.N., Lange,S.C., Myers,R.S., Davisson,V.J. and Smith,J.L. (2003) Toward understanding the mechanism of the complex cyclization reaction catalyzed by imidazole glycerolphosphate synthase: crystal structures of a ternary complex and the free enzyme. *Biochemistry*, **42**, 7003–7012.
33. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.
34. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
35. Shatsky,M., Nussinov,R. and Wolfson,H.J. (2004) A method for simultaneous alignment of multiple protein structures. *Proteins*, **56**, 143–156.
36. Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
37. Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–380.
38. Camon,E., Barrell,D., Lee,V., Dimmer,E. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database—an integrated resource of GO annotations to the UniProt Knowledgebase. *In Silico Biol.*, **4**, 5–6.